Dynamics of an insertion sequence infection in a spatially structured environment

Bichsel, Manuel ; Barbour, A D ; Wagner, Andreas

Abstract: Bacterial insertion sequences (ISs), the simplest form of autonomous mobile DNA, depend on their prokaryote hosts to spread in a spatially structured environment. We use a spatially explicit metapopulation model to simulate the spread of an IS that can have both detrimental and beneficial effects on its host cell. We find that, on the one hand, the spatial structure of the metapopulation and cell dispersal between subpopulations have no strong effect on the time to full infection of the metapopulation. On the other hand, factors that influence the IS infection dynamics within a subpopulation have a strong effect on that time. These factors are mainly the fitness benefit of an IS and the rate of horizontal gene transfer. We also find that the infection process of a metapopulation is very erratic in its early phase. Finally, we show that the infection's success depends critically on the initially infected subpopulation.

DOI: https://doi.org/10.1142/s0218339018500079

The dynamics of an insertion sequence infection in a spatially structured environment

Manuel Bichsel *and A. D. Barbour †and Andreas Wagner ‡

Bacterial insertion sequences, the simplest form of autonomous mobile DNA, depend on their prokaryote hosts to spread in a spatially structured environment. We use a spatially explicit metapopulation model to simulate the spread of an insertion sequence that can have both detrimental and beneficial effects on its host cell. We find that, on the one hand, the spatial structure of the metapopulation and cell dispersal between subpopulations have no strong effect on the time to full infection of the metapopulation. On the other hand, factors that influence the IS infection dynamics within a subpopulation have a strong effect on that time. These factors are mainly the fitness benefit of an insertion sequence and the rate of horizontal gene transfer. We also find that the infection process of a metapopulation is very erratic in its early phase. Finally, we show that the infection's success depends critically on the initially infected subpopulation.

Keywords: transposable element, infection dynamics, spatial model, simulation

1. Introduction

Mobile DNA has been fascinating researchers since its discovery in the 1940s by Barbara McClintock.[1] Why does it persist, even though its effects are detrimental to its host cells on average? The persistence of mobile DNA is especially puzzling in prokaryotes. While even detrimental mobile DNA may spread in a sexually reproducing eukaryote, especially if the mobile DNA's effects are recessive,[2–4] the detrimental effects of mobile DNA cannot be masked in this way in an asexually reproducing prokaryote. In addition, due to the generally high effective population size of prokaryotes, even small detrimental fitness effects of mobile DNA may cause strong negative selection. The spread of mobile DNA in prokaryotes is thus more difficult to explain.

In this paper, we study the spread of insertion sequences (ISs), the simplest form of autonomous mobile DNA, in spatially structured metapopulations of prokaryotes. ISs are short DNA sequences (0.7–2.5 kb) that typically encode only one enzyme, transposase, which enables transposition. During transposition, an IS usually gets excised and inserted into another location in the genome (conservative transposition), but occasionally the IS is copied during the transposition process (replicative transposition).[5] While the number of active IS copies in a genome increases through

*Institute of Evolutionary Biology and Environmental Studies, University of Zürich, Switzerland
†Institute of Mathematics, University of Zürich, Switzerland
‡Institute of Evolutionary Biology and Environmental Studies, University of Zürich, Switzerland
(corresponding author; +41 (0)44 635 61 41, andreas.wagner@ieu.uzh.ch)

replicative transposition, it decreases through IS excision and mutations that render transposase ineffective. Insertion sequences are vertically transmitted through inheritance, but ISs can also be horizontally transmitted by horizontal gene transfer (HGT) between prokaryotes.[6]

ISs in general have a detrimental fitness effect on their host cell, not only due to their transposition activity, but also because of genome rearrangements that can occur if a genome contains more than one IS copy.[7–9] Occasionally, though, ISs benefit their hosts. Many ISs contain an outwardly directed, entire or partial promoter that can increase the expression of a nearby gene.[10] Furthermore, two synchronously transposing ISs can mobilize genes lying between them. These genes often confer resistance to antibiotics, encode toxins, or allow for new metabolic functions.[11–13] The composite transposon (also called compound transposon) that has thus been created can then insert into a plasmid and spread through a host cell population. A still unresolved question is whether ISs persist because they are occasionally beneficial to their hosts[9,14,15] or because HGT is strong enough to overcome their detrimental fitness effects.[3,16–19] In earlier work, we showed that a purely detrimental IS infection can successfully invade an uninfected host cell population, provided that the HGT rate exceeds the detrimental effect of ISs on a host cell.[20] For a specific IS family, we also estimated the HGT rate that would be needed to reach the distribution of IS copies per genome which can be observed in the wild. We showed that this HGT rate is well within the range of HGT rates estimated by experiments, but that the infection process would take an unrealistically long time if it depended only on HGT.[21] We then concluded that beneficial effects of an IS infection on infected cells, although they may be temporary, can play an important role in speeding up the infection process. This is in accordance to an earlier finding of one of us[22] who has shown that the sequence divergence of IS copies within genomes is much lower than between genomes, indicating that ISs might undergo 'burst and bust' cycles of infection and extinction in local populations which may take several hundred thousand host cell generations.

Many prokaryotes live in a spatially structured environment, which may influence the dynamics of an IS infection. A few studies explore the infection dynamics of mobile DNA in spatially structured metapopulations of multicellular organisms.[23,24] However, to our knowledge, there exists no analysis of prokaryotic IS infection dynamics in a spatially structured environment, even though theoretical predictions for single populations without spatial structure exist.[20,25,26] In such a spatially structured environment, ISs may spread to new prokaryote host populations through the dispersal of infected cells.

While the Baas Becking hypothesis for microorganisms, "*Everything is everywhere, but the environment selects*" (Ref. 27, translated), is still debated among microbiologists (for a review, see Ref. 28), it has been shown that microorganisms can indeed spread widely. This holds especially for spore-forming prokaryotes, which can spread not only through migrating host animals,[29] but also directly through the air by wind[30,31] or through the oceans by water currents[32,33] over long (even inter-

continental) distances. Many prokaryote species may thus have a global distribution, although other prokaryote species, e.g. the extremophiles, seem to show a more restricted, local distribution.[34] However, even in a globally distributed prokaryote host species, the spreading of an IS does not occur instantaneously, and IS infection dynamics may be affected by the spatial structure of the environment.

A large body of literature exists about spatial invasion processes in ecology[35,36] and about spatial infection processes in epidemiology (Ref. 37, chap. 7). Some phenomena that may be observed during invasion or infection processes in the wild are directly linked to space. One such phenomenon is the appearance of infection waves, where a geographical area containing infected individuals or subpopulations shows a well-defined, expanding front line. Another relevant phenomenon is the extinction/rescue effect, whereby an invading species may become temporarily extinct in a specific habitat patch (subpopulation), only to be rescued by immigrating individuals from another patch. In its extreme form, this effect can lead to a source-sink dynamics, where an invading species only persists in a specific patch because of constant immigration from other patches.[36] Phenomena like these can only be analyzed with spatially explicit models, where subpopulations or even individuals occupy specific spatial locations. With such a model, one can then explore the effect of different spatial distributions of subpopulations on the speed of invasion or infection, and examine processes of pattern-formation during invasion or infection.

Here we model a metapopulation consisting of spatially separate subpopulations, where all subpopulations initially contain only uninfected cells, with the exception of one subpopulation, which in addition contains a few infected cells. The infection dynamics of ISs in the metapopulation is determined by *local* processes within each subpopulation, such as HGT and the competition between infected and uninfected host cells, and by *global* processes between subpopulations, such as cell dispersal. We use stochastic and deterministic models to analyse the influence of local processes on the infection dynamics in a subpopulation. In addition, we use a spatially explicit, stochastic metapopulation model to simulate the spreading of an IS infection. Using this latter model, we address two main questions concerning beneficial and detrimental IS infections in a spatially structured environment. First, how do spatiality and local or global processes influence IS infection speed? And second, what is the role of the initially infected subpopulation in the infection process?

2. Model and Methods

In stochastic population modeling, two sources of random variation in population dynamics can be distinguished.[38–41] The first is demographic stochasticity due to the randomness of birth and death events in a finite population. The second is environmental stochasticity due to random variation of birth and death rates over time and space. The effects of demographic stochasticity are strongest for small populations and dominate during the initial phase of population growth. In contrast, the effects of environmental stochasticity dominate during later phases, when popula-

tion sizes are already large. It is very difficult to assess the extent of environmental stochasticity for bacterial host cells over long time periods and over large spatial scales. Rather than assuming some arbitrary variation in cell division rate and death rate over time and space, we therefore only take into account demographic stochasticity in our simulation model. The effects of neglecting environmental stochasticity are mitigated by the fact that our unit of time is a host cell generation, defined by the mean time between cell divisions. As long as model rates do not change in relation to the rate of cell division over time and space, times measured in generations will therefore not change either (although absolute times may change).

In our modeling, we distinguish between a subpopulation level and a metapopulation level. In the next two subsections, we describe both levels in more detail.

2.1. *Subpopulation level*

We assume that subpopulations have no spatial structure but are well-mixed. We use a stochastic simulation model and a combination of a branching process model with a deterministic model to understand IS dynamics on the subpopulation level. The stochastic simulation model (panel A in figure 1) is embedded into the stochastic simulation model of the metapopulation, and together they allow us to determine the influence of both local and global processes on the infection dynamics of a metapopulation in a spatially structured environment. Both the branching process model (panel B in figure 1) and the deterministic model (panel C in figure 1) closely follow the design of the stochastic simulation model. Combined, the branching process model and the deterministic model allow us to determine the influence of local processes on the infection dynamics in a subpopulation. The next paragraph describes the assumptions we make to keep our subpopulation models tractable.

We assume that each host cell can carry at most a single IS in its genome. This is not a strong limitation, as the IS count distribution of all IS families in the wild is strongly L-shaped, i.e. for any IS family, most genomes contain no IS copy, many genomes contain one copy, and only few contain more than five copies.[21,22,42] To reduce model complexity, and especially to increase simulation speed, it is therefore reasonable to allow for at most one IS copy per genome. In all our models, an IS insertion can either have a detrimental effect $s_d < 0$ or a beneficial effect $s_b > 0$ on host cell fitness. Empirical data show that the fitness effect of an IS depends on the location of the insertion in the genome. In a few locations, an IS may have a beneficial fitness effect, for example by promoting the expression of a nearby gene.[10] In most other locations, the same IS may have a detrimental fitness effect, for example by inserting into a gene, thus silencing the gene.[43] Noncoding regions, into which an IS could safely insert, constitute only a small fraction of about 10% of prokaryotic genomes,[44] and even noncoding regions include many regulatory DNA sequences that are sensitive to disruption. Genome locations with beneficial side effects are therefore rare in comparison with genome locations with detrimental side effects. Keeping these observations in mind, we assume that the conversion from a detri-

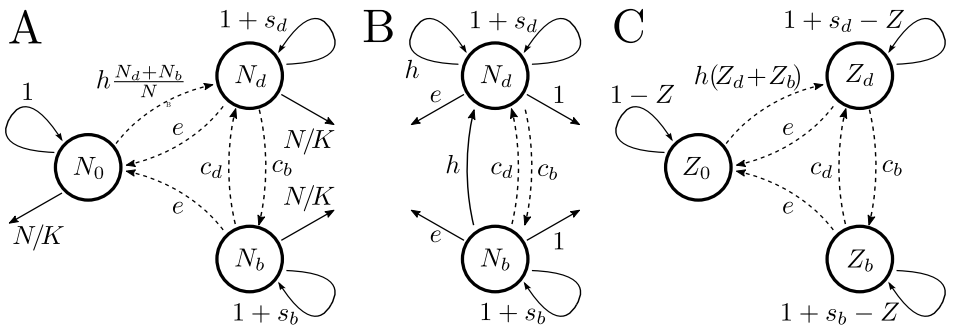

Fig. 1. *Subpopulation model design.* In the stochastic simulation model (panel A) and in the branching process model (panel B), N_0, N_d and N_b are the cell counts of uninfected, detrimentally infected and beneficially infected cells, respectively, and K is the carrying capacity. $N = N_0+N_d+N_b$ is the total cell count. Observe that uninfected cells (N_0) are not included in our branching process model. In the deterministic model (panel C), $Z_0 = N_0/K$, $Z_d = N_d/K$ and $Z_b = N_b/K$ are normalised population sizes, where K is the carrying capacity. $Z = Z_0 + Z_d + Z_b$ is the total normalised population. For all models, $s_d < 0$ and $s_b > 0$ indicate the fitness cost of a detrimental infection and the fitness benefit of a beneficial infection, respectively. The parameter e denotes the IS excision rate, c_d and c_b are the conversion rates from beneficial to detrimental infection and from detrimental to beneficial infection, respectively, and h is the HGT rate. All rates are per cell and cell generation. Solid and dashed arrows indicate a change and no change in the total size of the modeled population, respectively.

mentally infected cell to a beneficially infected cell by conservative transposition occurs with a rate c_b that is 1000 times lower than the rate of conversion c_d from a beneficially infected cell to a detrimentally infected cell. We set the conversion rate c_d of beneficial to detrimental infection to the conservative transposition rate in the wild, because conservative transposition is the only way to switch between beneficial and detrimental infection in our model, and because transposition into a genome location with beneficial side effects is very rare. An IS can get excised and lost or suffer an inactivating mutation with rate e.[8,45] Furthermore, an IS can be copied from an infected host cell's genome to an uninfected host cell's genome by HGT. The process by which an IS can be transferred from an infected to an uninfected host cell's genome can be complex. For example, it may involve bacterial conjugation, followed by the transfer of an IS-bearing plasmid, and the IS's transposition into the host genome (into sites whose specificity may depend on the IS itself). To keep the model tractable, we collapse all this mechanistic complexity of HGT into a single parameter h that represents the HGT rate.[46] Because genome locations that lead to a beneficial infection are rare, we assume that all cells that acquire an IS by HGT will initially be detrimentally infected. Transposition, excision and HGT rates reported in the literature cover a wide range of values (table 3 in Appendix A). We therefore systematically vary the parameters of local processes over a wide range, starting from a default parameter set, with which we conduct most of our

analysis and simulations. Table 1 shows for each local parameter its default value and the parameter range that we explore.

Table 1. Default values and ranges of values of local parameters used in our simulation model.

Symbol	Description	Default value	Range of values
s_d	detrimental fitness effect	-10^{-4}	$[-10^{-3}, -10^{-5}]$
s_b	beneficial fitness effect	10^{-4}	$[10^{-7}, 10^{-1}]$
c_d	detrimental conversion rate	10^{-6}	$[0, 10^{-6}]$
c_b	beneficial conversion rate	10^{-9}	$[0, 10^{-9}]$
e	excision rate	10^{-9}	$[0, 10^{-9}]$
h	HGT rate	10^{-4}	$[10^{-7}, 10^{-1}]$

Note: All rates are indicated as numbers of events per cell and generation.

We now discuss the specifics of our stochastic simulation model (panel A in figure 1). The effective size N_e of prokaryote populations usually exceeds 10^8 individuals.[47] We therefore assume in our stochastic simulation model that each subpopulation has a carrying capacity of $K = 10^9$ host cells (computational limitations prevent us from exploring much larger populations). At the start of the simulation, all subpopulations contain 10^9 uninfected cells, and one subpopulation – the initially infected subpopulation – usually contains an additional 100 beneficially infected cells. Ideally, we would want the simulation to start with one infected cell, so as to simulate the spread of an IS that has been generated de novo or has been reactivated by mutation from an inactive state. However, we choose an initial number of 100 infected cells as a compromise between two conflicting requirements. On the one hand, as the number of initially infected cells increases, so does the probability that a metapopulation becomes fully infected during simulation. Because we are mainly interested in metapopulations that become fully infected, we need fewer simulations if we increase the number of initially infected cells. On the other hand, an increase in the number of initially infected cells decreases the time for a metapopulation to become fully infected and, more importantly, also decreases the variation in that time. By substantially increasing the number of initially infected cells, we would thus obtain unreliable simulation results, especially regarding the variability of the infection process. In exploratory simulations starting with one beneficially infected cell, the median time needed to reach 100 infected cells was 60 generations for those simulations that did reach 100 infected cells, with an interquartile range of 45 generations. This interquartile range is very small compared with the interquartile range in the time to full infection of a metapopulation, which typically is of the order of 10 000 generations. It is therefore reasonable to choose an initial number of 100 beneficially infected cells instead of just one infected cell, thus strongly decreasing the number of simulations needed to obtain a statistically meaningful sample size of fully infected metapopulations. Nevertheless, for comparison with the results obtained from the

deterministic model, we also perform simulations of a subpopulation starting with 10^9 uninfected cells and only one beneficially infected cell, using the set of default parameters for local processes described in table 1.

In our stochastic simulation model, we simulate the IS infection dynamics in each subpopulation by using the tau-leaping algorithm.[48,49] We use this algorithm to calculate the length of each time step in the simulation and to determine the numbers N_0 of uninfected cells, N_d of detrimentally infected cells, and N_b of beneficially infected cells at the end of each time step. For a more detailed description of the tau-leaping algorithm, see Appendix B.

We now turn to a discussion of our combination of a branching process model with a deterministic model, where we use a branching process for the early phase of an IS infection inside a subpopulation and model the later phase deterministically.[50] For the parameter range of local processes described in table 1, we use a multi-type, continuous-time Markov branching process (panel B in figure 1) to calculate the probability of infection persistence, of the time needed to reach a certain threshold of infected cells, and of the distribution of detrimentally and beneficially infected cells when reaching this threshold. After some experimentation we chose a threshold of 10^6 for the total of detrimentally or beneficially infected cells. This threshold is on the one hand big enough for the asymptotic results from the theory of branching processes to apply (e.g. the distribution of the time to reach the threshold, or the distribution of detrimentally and beneficially infected cells). On the other hand, it is small enough for the proportion of infected cells in a subpopulation with $K = 10^9$ to be negligible. This is important because branching processes in the long term either die out or grow infinitely, without self-regulation. In Appendix C, we give a more detailed description of our branching process model. There we also show how the multi-type branching process model can under some circumstances be simplified to a single-type, birth-and-death process model. For this simpler model we analytically deduce the survival probability $p_{\mathrm{surv}} \approx s_b$ of an IS infection if $h \leq |s_d|$, i.e. if HGT cannot overcome the fitness cost of a detrimental IS.

We use a deterministic model of a subpopulation (panel C in figure 1) to analyse the later phase of an IS infection inside a subpopulation. The deterministic model's initial conditions are based on results from our branching process model, namely the median time to reach the threshold of 10^6 infected cells (if the infection persists), and the distribution of detrimentally and beneficially infected cells at that time. We formulate the deterministic model in terms of normalised population sizes $Z_0 = N_0/K$, $Z_d = N_d/K$ and $Z_b = N_b/K$. The infection dynamics in a subpopulation without immigrating or emigrating cells can then be described by the following system of ordinary differential equations (ODEs), where $Z = Z_0 + Z_d + Z_b$ is the total normalised population size:

$$\begin{aligned}
\dot{Z}_0 &= (1 - Z)Z_0 - hZ_0\,(Z_d + Z_b) + e\,(Z_d + Z_b) \\
\dot{Z}_d &= (1 + s_d - Z)Z_d + hZ_0\,(Z_d + Z_b) + c_dZ_b - c_bZ_d - eZ_d \\
\dot{Z}_b &= (1 + s_b - Z)Z_b - c_dZ_b + c_bZ_d - eZ_b
\end{aligned} \tag{2.1}$$

Starting with $Z_0 = 1$ (corresponding to $N_0 = 10^9$), and $Z_d = N_d/K$ and $Z_b = N_b/K$ based on the branching process model's prediction of N_d and N_b at the time when the infection threshold of 10^6 cells has been reached, we solve this system numerically, using the same parameters as in the branching process model.

2.2. *Metapopulation level*

Even though many prokaryotes are motile, they are slow and move only short distances on their own. However, prokaryotes can also be dispersed passively by wind, water and migrating animals over long distances. We study the influence of this passive dispersal on the IS infection dynamics using a metapopulation of spatially distributed subpopulations inside a circular region of 100 km radius, roughly the size of an ecoregion.[51] Using this radius, the region is large enough to possibly observe spatial phenomena like infection waves, while being small enough to keep the number of subpopulations low and allow for manageable simulation times.

For most of our simulations, we cover the circular region with a hexagonal lattice in such a way that a vertex of the lattice is located at the center of the region. We then place a subpopulation at each vertex. As to the distance between neighbouring subpopulations (i.e. the edge length of the lattice), we seek a compromise between a high level of detail in our simulations (many subpopulations, and therefore short distances between them) and simulation feasibility (few subpopulations, and therefore long distances between them). After exploratory simulations with different distances between neighbouring subpopulations, we chose a value of $100/8 = 12.5$ km, which results in a default metapopulation comprising a total of 241 subpopulations (panel A in figure 2). Simulations with halved and doubled distances showed that our choice of edge length does not influence our results. In order to assess the effect of spatial metapopulation organisation on the infection dynamics of ISs, we use two additional types of spatial subpopulation distribution. In the first, the 241 subpopulations show a uniform random distribution (panel B in figure 2). In the second, the 241 subpopulations show a clustered distribution, determined by a mixture of habitable and uninhabitable landscape types with a contagion index of 0.4 (panel C in figure 2). The contagion index[52,53] is a measure of the clumpiness of a landscape, where a value of zero signifies no clumping, i.e. a thorough mixture of landscape types, and a value of one signifies that the landscape consists of only one type. For comparison purposes, we also conduct simulations with a spatially unstructured, single population with the same initial number of cells as the spatially structured metapopulations ($241 \cdot 10^9$ uninfected cells plus 100 beneficially infected cells).

We simulate cell dispersal by exchanging cells between all subpopulations every 10 generations. We used exploratory simulations in which subpopulations exchange cells every 2 and 50 generations, respectively, to ensure that our choice of the time between cell exchange would not influence our results. We compute the number of cells that migrate from one subpopulation to another based on a Poisson distribution. The distribution's mean is computed using a cell dispersal rate function that

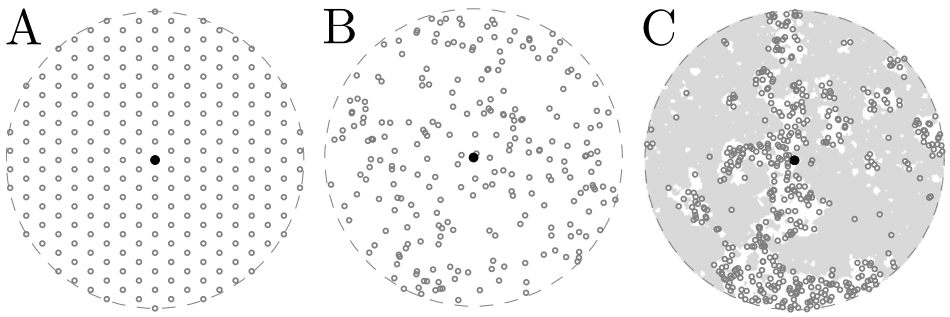

Fig. 2. *Metapopulation model design.* Regular (panel A), uniform random (panel B),and clustered (panel C) spatial distributions of 241 subpopulations inside a circular region with a radius of 100 km are shown; the initially infected subpopulation in the center is indicated as a black closed circle; the landscape in the clustered spatial distribution of subpopulations (panel C) has a contagion index of 0.4.

describes how the mean number r of cells migrating from one subpopulation to another during one generation depends on the distance d (in kilometers) between the two subpopulations. To our knowledge, there exist no direct measurements of prokaryotic cell dispersal in the wild. Our default rate function is derived from indirectly measured cell dispersal data from[54] and has the form of a power function, $r = r(d) = 239.6 \cdot (d + 0.1)^{-0.53}$ migrating cells per generation (subsection 3.2 and figure 4). In view of the uncertainty about dispersal rates over long distances, we also conduct simulations using rate functions with different proportions of dispersal over short and long distances, while keeping the mean dispersal rates identical to those in a metapopulation with regularly distributed subpopulations on the vertices of a hexagonal grid (panel A in figure 2), to allow for fair comparisons between rate functions. To explore the effects of steeper power functions with exponents smaller than -2, which have been estimated by other authors for pollen and plant seeds,[55,56] we use an exponent of $4 \cdot (-0.53) = -2.13$ and readjust the multiplicative constant in our default rate function so that the mean dispersal rate in a metapopulation with regularly distributed subpopulations is the same as for the default rate function. This leads to the rate function $r = 91\,728.4 \cdot (d+0.1)^{-2.13}$ migrating cells per generation. We also include two extreme dispersal rate functions: a constant function, where the dispersal rate does not depend on the distance between two subpopulations, and a nearest neighbour function, where cell dispersal occurs only between immediately neighbouring subpopulations, i.e. there exists a threshold distance below which dispersal is constant and above which dispersal is impossible. We again choose the constants so that the mean dispersal rate in a metapopulation with regularly distributed subpopulations is the same as for the default rate function. This leads to the constant rate function $r = 24.5$ migrating cells per generation for all subpopulations. For the nearest neighbour rate function,

the cell dispersal rate to the nearest six subpopulations on the hexagonal grid is $r = 1065.3$ migrating cells per generation, and $r = 0$ to all other subpopulations.

We say that a subpopulation has reached *complete infection* if it has an infection prevalence of at least 95%, i.e. if at least 95% of all cells are infected. If at least 99% of all subpopulations are completely infected, we say that the metapopulation has reached *full infection*. This definition ensures that a few subpopulations with low cell dispersal rates will not unduly distort our results. We then use the time (in cell generations) to full infection as a simple indicator of the infection dynamics. For the large, spatially unstructured population with a carrying capacity of $241 \cdot 10^9$ cells, we correspondingly define the time to full infection as the time to reach an infection prevalence of at least $0.95 \cdot 0.99 = 94.05\%$.

For metapopulations with subpopulations on a hexagonal lattice, we conducted 5000 simulations per metapopulation, of which typically about 50 simulations led to full infection. For metapopulations with randomly distributed subpopulations, we conducted 100 simulations on each of 50 different realisations of a random subpopulation distribution. All times are given in cell generations. The time resolution of the output for metapopulations is 500 generations. This means that every 500 generations, the number of uninfected, detrimentally infected and beneficially infected cells existing in the subpopulations is reported, as is the number of immigrating or emigrating cells for every subpopulation during the previous 500 generations.

For comparison with our combination of a branching process model with a deterministic model of a subpopulation, we also conducted 10^5 simulations of a single subpopulation, starting with 10^9 uninfected cells and one beneficially infected cell, using a time resolution of 10 generations. This led to 14 complete infections.

We wrote the simulation program in C++ (gcc version 4.6.4), using the Boost libraries (version 1.49.0). The documented source code is available upon request. We analysed the deterministic model for a subpopulation and all simulation data using Mathematica (versions 10 and 11).

3. Results

Whether an infection of the metapopulation will succeed or not depends mainly on the fate of the infection in the initially infected subpopulation, because it is unlikely that during the early phase of an infection, any of the few infected cells in this subpopulation will be dispersed to other, still uninfected subpopulations. In a first step, we therefore analysed the infection dynamics and its dependence on local processes in a single subpopulation (next subsection). We then widened our scope, investigated bacterial cell dispersal based on existing data and examined the influence of both local and global processes on the infection dynamics in the metapopulation (later subsections).

3.1. *In a subpopulation, the fitness benefit of an IS and HGT strongly influence infection speed, but IS excision and conversion do not*

We used both a stochastic simulation model and a combination of a branching process model with a deterministic model to analyse the IS infection dynamics in a single, well-mixed subpopulation ($K = 10^9$) without any spatial structure. Figure 3 shows a comparison of the infection dynamics between the two approaches. In both cases, we used default values for local process parameters (table 1) and chose complete infection as the endpoint of the infection process, i.e. the time to reach an infection prevalence of 95% in the subpopulation. In figure 3, we show the infection process of those three simulations whose times to complete infection were closest to the quartiles of that time in all 14 simulations (out of 10^5) that reached complete infection. While the IS infection process of a subpopulation is very erratic during the early phase, when there are only few infected cells, the infection process becomes more deterministic later on, when a large fraction of cells is infected. To compare the infection dynamics of the stochastic simulation model with the dynamics of the combined model, we therefore align in figure 3 the time lines of all infection processes at the time point of complete infection, i.e. we set the time point at which complete infection has been reached to $t = 0$, and show all other time points during the infection process as negative values.

As can be seen in figure 3, the combination of a branching process model with a deterministic model predicts the infection dynamics of the stochastic simulation model well, at least for the default values for local process parameters that we used. It can also be seen that the population of infected cells initially seems to grow faster than during the later infection phase. This is a well-known feature of stochastic processes (e.g. Ref. 57, p. 12, and Ref. 58, p. 154), here caused by the following dynamics during the early phase of an IS infection: the risk of infection extinction through random events is very high when the number of infected cells is still low, and only those infections which by chance quickly grow to a substantial number of infected cells (i.e. 10^3 to 10^4 cells in figure 3) evade this risk and may persist, thus shortening the expected time to reach a 95% infection prevalence.

We note that the small number of 14 simulations that have reached complete infection out of 10^5 simulations that have been conducted is expected, as we now show. According to equation (C.7) in Appendix C, the probability of a simulation starting with one beneficially infected cell and with default parameters, where $h \leq |s_d|$, is $p_{\text{surv}} \approx s_b = 10^{-4}$. The number of simulations that reach complete infection therefore has a binomial distribution $\text{Bi}(n; p)$ with $n = 10^5$ and $p = s_b = 10^{-4}$. The expected number of simulations that reach complete infection is then $np = 10$, and the standard deviation of this number is $\sqrt{np(1-p)} \approx 3.2$.

Using a combination of a multi-type branching process model with a deterministic model, we analysed the influence of local processes on the infection dynamics in a subpopulation. To this effect, we systematically decreased and increased tenfold

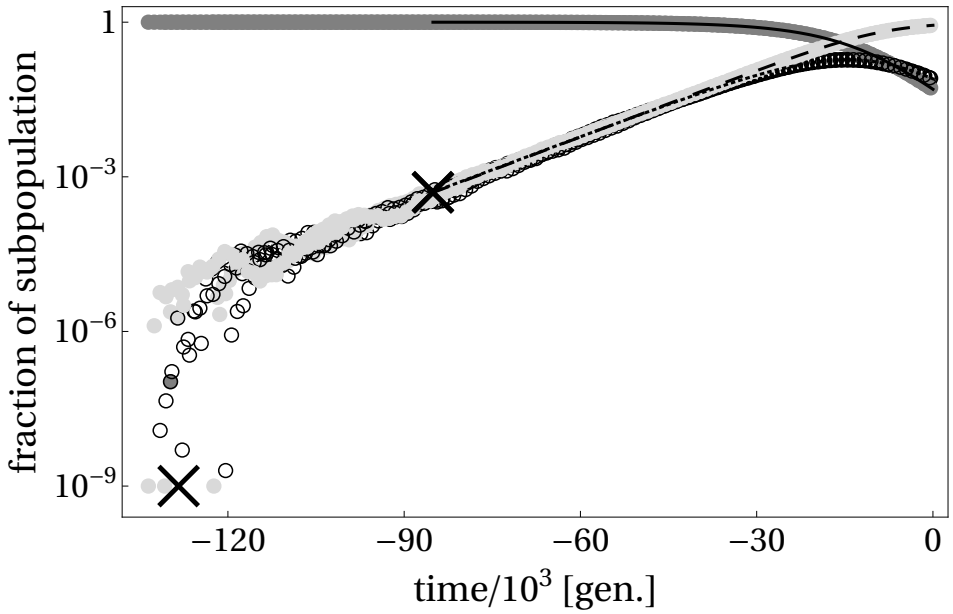

Fig. 3. *In a subpopulation, the combination of a branching process model with a deterministic model predicts the infection dynamics of the simulation model well.* Fraction of uninfected cells (dark grey disks, solid line), beneficially infected cells (light grey disks, crosses and dashed line) and detrimentally infected cells (circles, cross and dotted line) in a subpopulation over time. We determined these fractions by using the simulation model (disks and circles), and by numerically solving the combination of a branching process model (crosses) with a deterministic model (lines). In all models, we use the default parameter set. For the simulation model, the infection starts with a population of uninfected cells at carrying capacity ($N_0 = 10^9$), one beneficially infected cell ($N_b = 1$), and no detrimentally infected cells ($N_d = 0$). For illustration, we use the three simulations whose times to complete infection were closest to the quartiles of that time in all 14 simulations which reached complete infection. For the branching process model, the infection starts with one beneficially infected cell ($N_b = 1$) and no detrimentally infected cells ($N_d = 0$). We calculate the median time to reach a threshold of 10^6 infected cells and the numbers N_d and N_b of detrimentally and beneficially infected cells at that time. We then use N_d and N_b to construct the initial conditions for the deterministic model ($Z_d = N_d/10^9$, $Z_b = N_b/10^9$), together with $Z_0 = 1$, and calculate the solution of the ODEs based on the deterministic model. The horizontal axis indicates the time between infection start and complete infection with negative values, because we aligned the time lines of the combined branching process and deterministic model's solution and of all three simulations at the time point when complete infection was reached.

the fitness cost s_d, the fitness benefit s_b, and the HGT rate h, while keeping all other parameters constant. In addition, we analysed a model without IS excision and IS conversion ($e = c_d = c_b = 0$). Table 2 shows for different ratios of our model parameters the corresponding ratio in the time to complete infection.

Table 2 shows that IS excision and conversion do not influence the time to complete infection noticeably. However, a tenfold decrease in the fitness cost increases

Table 2. Dependence of the time to complete infection on model parameters.

param. γ	$\gamma_{\text{new}}/\gamma_{\text{default}}$	$t_{\text{new}}/t_{\text{default}}$
s_d	0.1	1.72
	10	1.02
s_b	0.1	4.9
	10	0.13
h	0.1	1.14
	10	0.15
e, c_d, c_b	0	0.99

Note: Ratios $t_{\text{new}}/t_{\text{default}}$ of the time t to complete infection for different ratios $\gamma_{\text{new}}/\gamma_{\text{default}}$ of model parameters γ using the combination of a branching process model and a a deterministic model inside a subpopulation. γ_{default} and t_{default} are the default value of the parameter and the time to complete infection for this default value, respectively. γ_{new} and t_{new} are the new value of the parameter and the time to complete infection for this new value, respectively. The default values of our model parameters are $s_d = -10^{-4}$, $s_b = 10^{-4}$, $h = 10^{-4}$, $e = 10^{-9}$, $c_d = 10^{-6}$, $c_b = 10^{-9}$ (all rates are indicated per cell and generation), and the corresponding time to complete infection is $t = 1.28 \cdot 10^5$ generations.

the time to complete infection by 72%. This apparent paradox is caused by the effect of a lower fitness cost on the distribution of detrimentally and beneficially infected cells: if the fitness cost is smaller than the HGT rate (here: $|s_d| = 10^{-5} < 10^{-4} = h$), even a population of purely detrimentally infected cells may persist and grow, albeit slowly. Therefore, for all infections that persist, the median time to reach complete infection is longer than if the fitness cost is equal to the HGT rate, and infections that persist contain more beneficially infected cells. The fitness benefit of a beneficial infection influences the time to complete infection even more than the fitness cost of a detrimental infection. As expected, increasing the fitness benefit decreases the time to complete infection, and decreasing the fitness benefit increases that time. For the HGT rate the situation is more complex in that the influence of the HGT rate h on the time to complete infection depends on whether the HGT rate is larger or smaller than the absolute value $|s_d| = 10^{-4}$ of the fitness cost in a detrimental infection. If the HGT rate is larger than this value, the infection time decreases with increasing HGT rate, but if the HGT rate is smaller, the time to complete infection does not greatly increase with decreasing HGT rate. This asymmetry arises because IS insertion is much more likely to be detrimental than beneficial, and we therefore assume that HGT leads to only detrimentally infected cells. To contribute

to the spread of an IS in a subpopulation, HGT then has to overcome the fitness cost $s_d = -10^{-4}$ of a detrimental IS. If the HGT rate is smaller than $|s_d|$, HGT cannot do so, and the infection will be driven by beneficially infected cells only, i.e. the time to complete infection does not depend on the HGT rate. But if the HGT rate is larger than $|s_d|$, HGT contributes in the same way to the infection dynamics as the fitness benefit s_b of a beneficial infection does.

3.2. *The dispersal kernel of bacterial cells is fat-tailed*

Having determined the influence of different local processes on the infection dynamics in a subpopulation, we then turned to the infection dynamics on a metapopulation level. We first searched the literature for information about the dispersal of bacterial cells in the wild.

While there is evidence of widespread, large-distance cell dispersal, e.g. through the air[30,31] or through the oceans,[32,33] reliable rates of cell dispersal are difficult to obtain. We based our default rate function on data that two authors[54] obtained by applying a cladistic method[59] to *Bacillus subtilis* and *Bacillus mojavensis* nucleotide sequences from three continents. Briefly, they treat the geographic location of each sequence as a multistate character and add the location to the sequence phylogeny. Then they derive the minimal number of migration events that are necessary to obtain the observed distribution of multistate characters over the phylogeny. From this number of migration events, they calculate the average number Nm of migrating individuals between subpopulations per generation (N is the subpopulation size, and m is the migration rate per individual and generation). The authors report Nm for three different distances d (in kilometers) between subpopulations: $(d, Nm) \in \{(30, 50), (500, 5.5), (10\,000, 2.25)\}$.[54] To obtain a dispersal rate function relating the mean rate r of migrating cells per generation between two subpopulations to the distance d between the subpopulations, we fitted a power law function of the form $r = r(d) = a \cdot (d + 0.1)^b$ to the three data points mentioned above. We used $d + 0.1$ instead of d in the function to avoid a singularity at $d = 0$. Using a power law function resulted in a better fit (residual sum of squares $RSS = 127.3$) than using other functions, e.g. an exponential function ($RSS = 1145.5$). This is in agreement with observations of long-distance dispersal of (small) pollen and plant seeds, where the dispersal kernel usually has a fat tail, i.e. the tail drops off slower than in an exponential function.[55,56,60] The best-fitting power law function we obtained was $r = 239.6 \cdot (d + 0.1)^{-0.53}$. It is shown in figure 4, together with the best-fitting exponential function and three other dispersal kernels that we used in our simulations (steep power law, constant and nearest neigbour function).

3.3. *Early on, an IS infection of a metapopulation is an erratic process*

Using the power law dispersal kernel based on data from the literature, we next analysed the early phase of an IS infection in a metapopulation, when infected cells

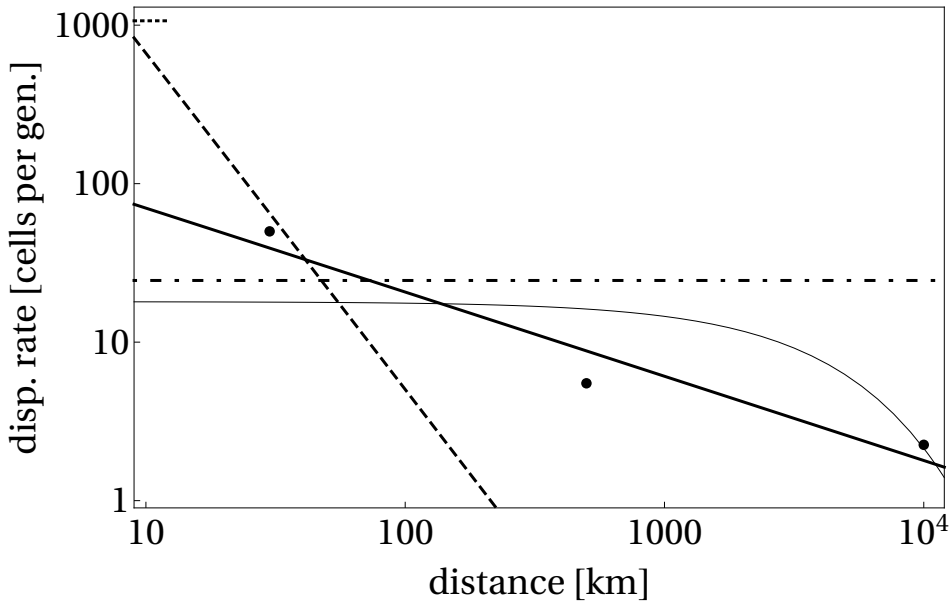

Fig. 4. *The dispersal kernel of bacterial cells is fat-tailed.* Dispersal rate of bacterial cells as a function of the dispersal distance. Shown are three indirect measurements (closed circles) from the literature,[54] together with several dispersal kernels $r = r(d)$, where r is the dispersal rate (cells per generation) and d is the distance in kilometers. The best-fitting power law function is $r(d) = 239.6 \cdot (d+0.1)^{-0.53}$ (solid, thick line). The best-fitting exponential function is $r(d) = 18.0 \cdot e^{-0.0021 \cdot d}$ (solid, thin line). In addition to the best-fitting power law function, we also used three other dispersal kernels in our simulations: a steep power law function $r(d) = 91\,728.4 \cdot (d+0.1)^{-2.13}$ (dashed line), a constant function $r(d) \equiv 24.5$ (dot-dashed line), and a nearest-neighbour function with $r(d) = 1065.3$ for $d \leq 12.5$ and $r(d) = 0$ for $d > 12.5$ (dotted line, only visible in top left corner), where $d = 12.5$ km is the distance between two subpopulations on the hexagonal grid. Note that both axes are logarithmic.

are restricted primarily to the initially infected subpopulation. Using default values for the local (table 1) and global parameters of the metapopulation, with subpopulations on a hexagonal grid (panel A in figure 2), only 49 out of 5000 simulations of our stochastic model led to full infection of the metapopulation. Moreover, only in one out of these 49 full infections of the metapopulation did the infection in the initially infected subpopulation die out temporarily and was rescued by another subpopulation, which it had infected before. This last observation illustrates the importance of the initially infected subpopulation for the infection process of the metapopulation. Due to the low number of infected cells in the initially infected subpopulation, the early phase of the infection is dominated by stochastic effects. The low number of 49 full infections of the metapopulation in 5000 simulations with default parameters is thus not unexpected, as can also be shown by the following calculation. According to equation (C.7) in Appendix C, the survival probability

of a subpopulation starting with one beneficially infected cell and with $h \leq |s_d|$ is $p_{\text{surv}} \approx s_b = 10^{-4}$. Because the 100 initially infected cells act independently of each other during the early phase of a subpopulation infection, when their number is low compared to the total number of host cells in the subpopulation, the number of persisting cell lineages, each starting with one beneficially infected cell, has a binomial distribution $\text{Bi}(n; p)$ with $n = 100$ and small $p = s_b = 10^{-4}$. The probability of an infection with 100 beneficially infected cells to persist and spread can then be approximated by a linearisation for small s_b as $1 - (1 - s_b)^{100} \approx 100 \cdot s_b = 0.01$. Because a fully infected subpopulation sooner or later inevitably leads to full infection of the metapopulation through cell dispersal, this is also the probability that a metapopulation gets fully infected. We thus expect about 50 out of 5000 simulations to lead to full infection of the metapopulation.

We call a subpopulation *successfully infected* if the number of infected cells in the subpopulation increases faster than the number of immigrating infected cells, i.e. if the infection in the subpopulation can spread on its own, without any immigration of infected cells. The median time for the initially infected subpopulation to successfully infect a second subpopulation is $2.0 \cdot 10^4$ generations, but with a large interquartile range of $1.8 \cdot 10^4$ generations.

A survival probability of $p_{\text{surv}} \approx s_b = 10^{-4}$ for an infection with default parameters that starts with one beneficially infected cell means that on average $1/p_{\text{surv}} = 10^4$ beneficially infected cells must immigrate into an uninfected subpopulation before an infection succeeds, i.e. most infections of a subpopulation die out.

Taken together, these observations demonstrate that the infection process is very erratic during its early phase, when both the number of infected subpopulations and the number of infected cells per subpopulation are low. Over time, when a larger fraction of cells in a subpopulation gets infected, the infection process becomes more deterministic. For example, the coefficient of variation of the time to reach an infection prevalence of 1% in the initially infected subpopulation (starting with a prevalence of 10^{-7}) is 15.0%, while the coefficient of variation of the time that passes between reaching an infection prevalence of 94% and of 95% in the initially infected subpopulation is only 5.7%.

3.4. *An IS infection of a metapopulation is not strongly slowed down by spatiality*

We next examined how the spatial distribution of subpopulations in a metapopulation influences the IS infection dynamics. Figure 5 shows the time to full infection for different spatial distributions, using default values for the local parameters of all metapopulations (table 1). The nonspatial metapopulation consists of a well-mixed single population starting with the same number of cells as the other metapopulations ($241 \cdot 10^9$ cells). The median time to full infection for this large, single population is $1.82 \cdot 10^5$ generations, and the median times to full infection for the

three metapopulations with a spatial structure of regional extent (median times between $1.97 \cdot 10^5$ and $1.98 \cdot 10^5$ generations) are at most 8.4% higher than the median time for the nonspatial population.

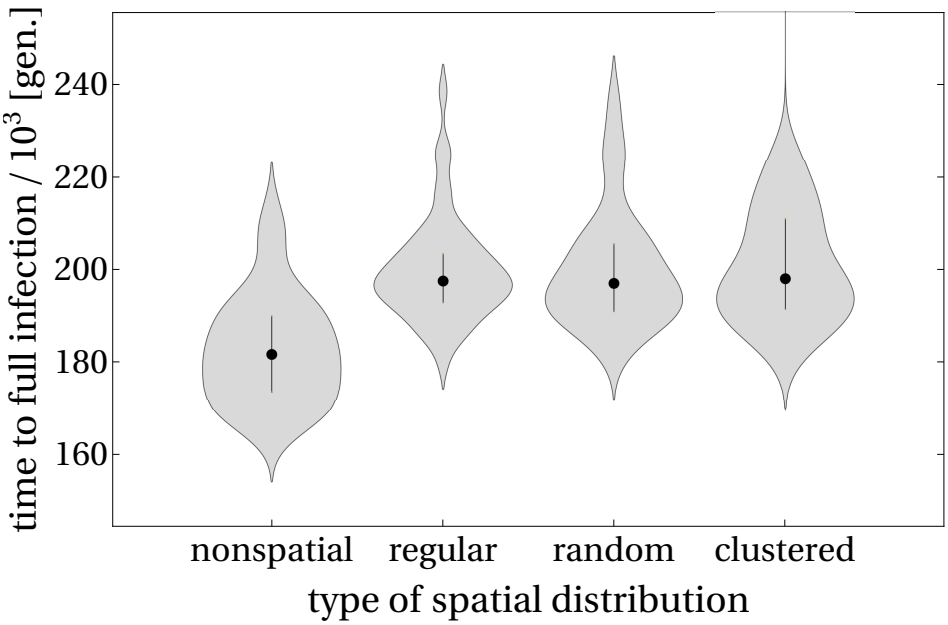

Fig. 5. *The spatial distribution of subpopulations influences IS infection dynamics only weakly.* Violin plot showing a kernel density estimate (shaded region), the median (dot) and the first and third quartile (endpoint of whiskers) of the time to full infection for a nonspatial, single population and the following three spatially structured metapopulations: regular (subpopulations on a hexagonal lattice), random (subpopulations uniformly distributed), clustered (subpopulations uniformly distributed in a clustered landscape). Number of observations (left to right): 62, 49, 57, and 63 simulations that led to fully infected metapopulations (out of 5000 simulations).

Figure 5 also shows that the type of spatial distribution (regular, random, or clustered) does not noticeably influence the median time to full infection. This may be a consequence of the fact that the default power law cell dispersal rate function, which is based on dispersal data from the wild, is relatively flat. For this function, the dispersal rate only drops from 70.3 cells per generation at a distance of 10 km between subpopulations to 20.9 cells per generation at a distance of 100 km between subpopulations. With this rate function, strongly varying distances between a subpopulation and its nearest neighbours, which are a characteristic of clustered landscapes, do therefore not pose any difficulties for the IS infection and will not slow it down noticeably.

3.5. *The shape of the dispersal function has only a limited influence on the infection speed in a metapopulation*

We also explored the effect of the cell dispersal rate function on IS infection dynamics. Figure 6 shows the time to full infection for four different cell dispersal rate functions: constant, default power law, steep power law, and nearest neighbour.

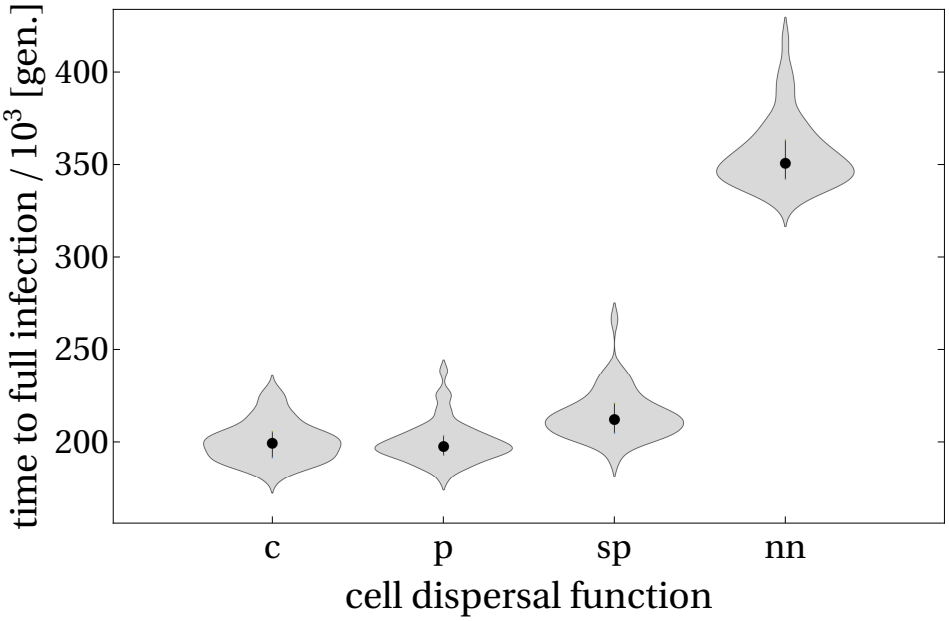

Fig. 6. *Realistic, data-based cell dispersal influences IS infection dynamics only weakly.* Violin plot showing a kernel density estimate (shaded region), the median (dot) and the first and third quartile (endpoint of whiskers) of the time to full infection for the constant (c), default power law (p), steep power law (sp) and nearest neighbour (nn) cell dispersal function. Number of observations (left to right): 58, 49, 48, and 51 simulations that led to fully infected metapopulations (out of 5000 simulations).

There is no noticeable difference in the median time to full infection for the constant rate function and the default power law rate function: the time for both is about $1.99 \cdot 10^5$ generations. Even the median times to full infection for the two power law rate functions differ by a mere 7.3%: $1.98 \cdot 10^5$ and $2.12 \cdot 10^5$ generations for the default and the steep power law rate function, respectively. Only the nearest neighbour dispersal rate function leads to a substantially longer median time to full infection than the other functions (e.g. 77.7% longer than for the default power law rate function). Taken together, figure 6 shows that the times to full infection for the default power law rate function, which is based on dispersal data from the wild, and two rate functions with flatter (constant) or steeper (steep power law)

shape are very similar. Even for the nearest neighbour rate function, the time to full infection has the same order of magnitude. For quite different shapes of the dispersal rate function, the infection speed does therefore not vary strongly inside a metapopulation with regional extent.

We examined the effect of changing the total cell dispersal rate while keeping the overall shape of the dispersal function the same. To this end, we used our default power law dispersal function $r(d) = c \cdot (d + 0.1)^{-0.53}$, where $c = 239.6$, and decreased or increased the multiplicative constant c by up to two orders of magnitude. Based on 5000 simulations for ten- and hundredfold smaller or larger values of c, we observed that such a decrease/increase of the dispersal rate leads to noticeably longer/shorter times to full infection than those obtained with the default rates. Specifically, while a tenfold reduced dispersal rate leads to a substantially increased time to full infection (+22.2%), a tenfold increased dispersal rate leads to a much smaller decrease in the time to full infection, from $1.98 \cdot 10^5$ to $1.84 \cdot 10^5$ generations (−7.1%). This is not unexpected, because the lower limit for the time to full infection is given by $1.82 \cdot 10^5$ generations, the time to full infection for a single, spatially unstructured population with carrying capacity $K = 241 \cdot 10^9$ cells. In summary, while the shape of the dispersal function, i.e. different proportions of dispersal over short and long distances, does not strongly influence the time to full infection, the total number of dispersed cells has a strong influence on this time.

3.6. *Both HGT rate and fitness benefit of an IS strongly influence infection speed*

We next examined how the fitness benefit s_b of an IS and the HGT rate h influence the infection dynamics. In contrast to the spatial distribution of subpopulations and the cell dispersal rate, these two parameters both reflect local processes within a subpopulation. In subsection 3.1 we have already shown that of all the local processes we consider in our models, the fitness benefit of an IS and HGT are the ones that have the strongest influence on the infection dynamics in a subpopulation. We therefore expect them to also have a strong effect on the infection dynamics in the metapopulation.

This is indeed what we observe in figure 7. For constant HGT rate $h = 10^{-4}$, the time to full infection increases with decreasing fitness benefit s_b of an IS (closed circles in the figure). At the same time, the extinction probability of an IS infection increases and for low fitness benefits $s_b < 10^{-5}$ reaches such high values that no infection persisted in 5000 simulations.

Similarly, for constant fitness benefit $s_b = 10^{-4}$, the time to full infection increases with decreasing HGT rate h (open circles in the figure), but only as long as the HGT rate is larger than the absolute value $|s_d| = 10^{-4}$ of the fitness cost in a detrimental infection. For even smaller values, $h < 10^{-4}$, HGT cannot overcome the fitness cost of a detrimental infection. The infection process is then driven by beneficially infected cells only, and the time to full infection thus becomes independent

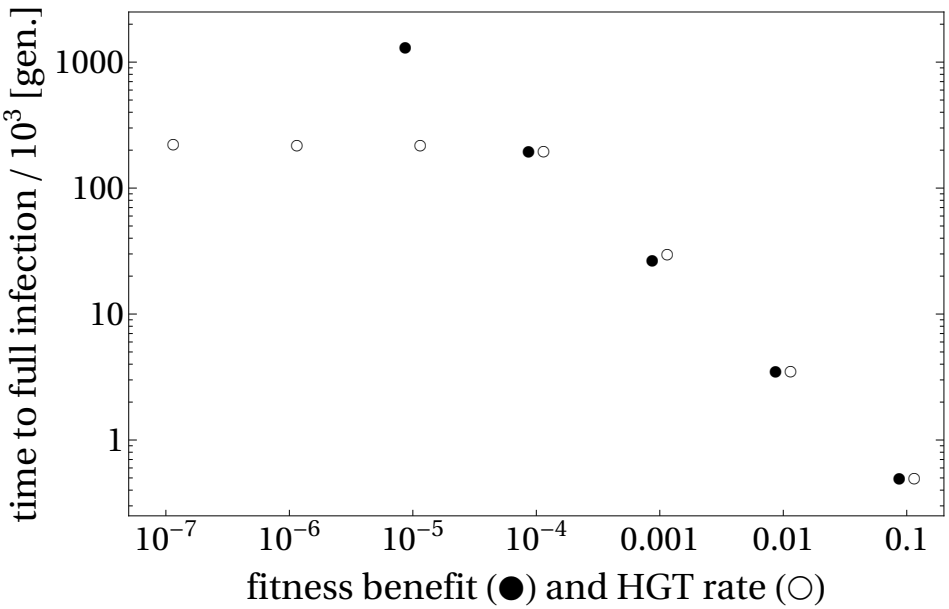

Fig. 7. *The fitness benefit of an IS and HGT rate influence IS infection dynamics in a metapopulation strongly.* Median of the time to full infection for different values of the fitness benefit s_b of an IS (closed circles) and of the HGT rate h (open circles). Circles have been slightly shifted horizontally to avoid overlap. The first and third quartiles are not shown because they are so close together as to be covered by the circles indicating the median. Number of observations for the time to full infection, with increasing fitness benefit: 0, 0, 7, 49, 509, 3168, 5000 simulations that led to fully infected metapopulations (out of 5000 simulations). Number of observations for the time to full infection, with increasing HGT rates: 46, 62, 53, 49, 470, 3080, 5000 simulations that led to fully infected metapopulations (out of 5000 simulations). Note that both axes are logarithmic.

of the HGT rate.

The HGT rate and the fitness benefit not only strongly influence the infection speed but also the probability that a metapopulation gets fully infected. This is reflected in the number of simulations that lead to fully infected metapopulations, which is reported in figure 7. Starting with the default parameter set, where $h = 10^{-4}$, $s_d = -10^{-4}$ and $s_b = 10^{-4}$, a tenfold change in the fitness benefit leads to a roughly tenfold change in the number of simulations that lead to full infection. This is to be expected, because equation (C.7) in Appendix C shows that for small values of the fitness benefit s_b and for the HGT rate equal to the absolute value of the fitness cost, $h = |s_d|$, the probability p of an IS infection first to spread through a subpopulation and then through the metapopulation is approximately equal to the fitness benefit, $p \approx s_b$. In contrast, figure 7 shows that a tenfold increase in the HGT rate increases the number of simulations that lead to full infection roughly tenfold, but a decrease in the HGT rate does not decrease that number. The reason

for this is the same as suggested above for the influence of the HGT rate on the time to full infection: only if the HGT rate is larger than the absolute value of the fitness cost, $h > |s_d|$, can HGT contribute in the same way to the infection as the fitness benefit s_b, and for smaller HGT rate, $h < |s_d|$, the persistence probability of the infection depends only on the fitness benefit.

3.7. *Metapopulation infection is mainly driven by the initially infected subpopulation*

To assess the role of the initially infected subpopulation during the infection process of a metapopulation, we focused on the 49 simulations with default parameters that led to full infection of the metapopulation. In 44 out of those 49 simulations, the initially infected subpopulation was the first to reach complete infection (i.e. reaching an infection prevalence of 95%). For each of these 44 simulations, we first sorted all 240 initially uninfected subpopulations according to their time to successful infection, thus creating a ranked list of subpopulations for each simulation (recall that we call a subpopulation successfully infected if its number of infected cells increases faster than the number of immigrating infected cells). Any given subpopulation need not have the same rank in all 44 ranked lists because the subpopulation may get successfully infected at different times during the 44 simulations we analysed. We then collected all the subpopulations of the same rank into groups, thus forming 240 rank groups with 44 subpopulations per group. For each of the 240 rank groups, the horizontal axis in figure 8 shows the median time to successful infection of the subpopulations in that rank group. The vertical axis shows, for the subpopulations in that rank group, the median fraction of immigrated, infected cells that originate from the initially infected subpopulation until the time when a subpopulation becomes successfully infected. In order to get an impression of the amount of variation in the data, the figure shows all three quartiles (dot for median, and endpoint of whiskers for first and third quartile) of the time to successful infection and of the fraction for groups number 1 (first successfully infected), 120, and 240 (last successfully infected).

Figure 8 shows that even for the last successfully infected subpopulation in each simulation (rank group number 240, with a median time to successful infection of $8.5{\cdot}10^4$ generations), a majority of immigrating, infected cells (median: 66.8%) originate from the initially infected subpopulation. This means that the initially infected subpopulation actually drives the infection process of the metapopulation, while subpopulations which are infected later on contribute much less to the infection of not yet infected subpopulations. We confirmed this observation with simulations in which cell dispersal was possible only from the initially infected subpopulation. Subpopulations that got infected later on were therefore not contributing to the infection of still uninfected subpopulations. Nevertheless, the time to full infection of the metapopulation increased by only 5.2%, compared with the original simulations in which dispersal is possible from all subpopulations.

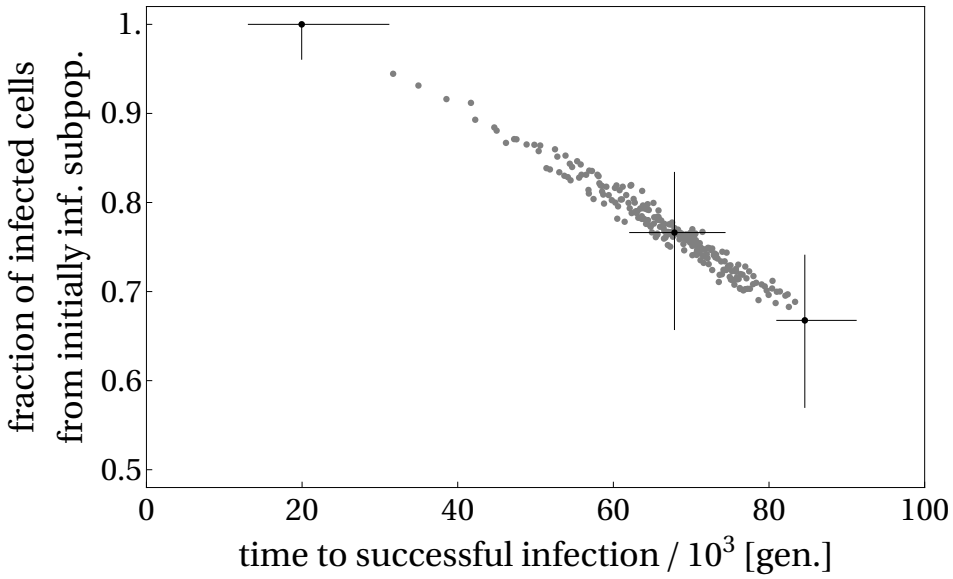

Fig. 8. *Metapopulation infection is mainly driven by the initially infected subpopulation.* Each gray dot corresponds to one of 240 "time rank" groups of initially uninfected subpopulations. The horizontal axis shows the median time to successful infection of a subpopulation, and the vertical axis shows the median fraction of cells that immigrate from the initially infected subpopulation until a subpopulation becomes successfully infected. In addition, the figure shows the median (black dot) and the first and third quartile (endpoint of black whiskers) of these quantities for three specific rank groups, i.e. numbers 1 (first successfully infected), 120 and 240 (last successfully infected). We constructed the time rank groups by first ranking all 240 initially uninfected subpopulations for each simulation by their time to successful infection and then collecting all subpopulations with the same rank into the same group.

While the emigration of infected cells allows the initially infected subpopulation to infect a whole metapopulation, the emigration (i.e. loss) of infected cells also takes a minor toll on the infection speed in the initially infected subpopulation. The median time to complete infection for the initially infected subpopulation is 10.9% longer than for a single population of the same size ($K = 10^9$ cells) without dispersal.

4. Discussion

We have shown in earlier articles that a detrimental IS can successfully invade a single host cell population if the HGT rate is larger than the detrimental fitness effect of the IS on the host cell, but that the infection process may take a very long time.[20,21] In this paper, we investigate the influence of spatiality on the spread of an IS through a host cell metapopulation by using a spatially explicit simulation model for the spatial infection dynamics of an IS that can be both detrimental

and beneficial to its host. In doing so, we also allow for beneficial effects of an IS on its host, which may shorten the infection process. Our key observations are that global processes at the metapopulation scale are less important for the infection dynamics than local processes within subpopulations, and that the initially infected subpopulation plays a key role in the infection of the metapopulation. We next discuss these observations in more detail.

4.1. *Global processes are less important for the IS infection dynamics than local processes*

In spatial ecology, one of the most important factors for the spread of an invasive agent is the dispersal kernel, a probability distribution describing the distance between a parent and its offspring (e.g. the probability distribution of the distance between a plant and one of its seedlings). The speed with which the agent spreads is mainly determined by the tail of the dispersal kernel.[61] If the kernel is fat-tailed (i.e. the probability of long distance dispersal diminishes slower than a negative exponential with distance), long-distance dispersal events usually lead to an infection that spreads without clearly defined and steadily expanding borders between infected and uninfected subpopulations. Instead, the initially infected subpopulation becomes surrounded by a fragmented patch of infected subpopulations, and many isolated, infected subpopulations far away from the initially infected subpopulation may exist. Those isolated subpopulations may then themselves become the seeds of fragmented patches of infected subpopulations, and the patches coalesce over time. In addition, the speed with which the infection spreads, defined by the square root of the infected area, divided by time, increases over time.[61–63] The spreading of an IS infection depends on the spreading of its prokaryote host, and we found that at least some prokaryotes have a fat-tailed dispersal kernel (figure 4), based on data from other authors.[54] We would therefore expect a patchy and irregular spreading of an IS infection in a metapopulation. This is indeed what we observe in figure 9 in Appendix D. This figure shows that the spreading of a simulated IS infection with a power law dispersal function proceeds irregularly, with many isolated infected subpopulations. We also conducted simulations which show that decreasing or increasing the spatial extent of a regional metapopulation has only a moderate effect on the time to full infection (figure 10 in Appendix E). Together with our observation that the spatial distribution of subpopulations does not limit the spreading of an IS inside the region we consider (figure 5), this suggests that ISs may also spread quickly over larger regions, even if the host cell habitat is clustered instead of uniform.

While the spreading of an IS is not strongly slowed down by spatiality and by dispersal, at least for the organisms and on the spatial scale we consider, IS infection speed depends very sensitively on processes within a host cell or between host cells in a local subpopulation (figure 7).

4.2. *An IS infection strongly depends on the initially infected subpopulation*

We show that the IS infection process depends critically on the initially infected subpopulation. The infection in that subpopulation must prevail against a high probability of dying out quickly, and the prevalence of infected cells needs to increase to high enough values so that the initially infected subpopulation has a substantial chance of infecting other subpopulations. We show that the infection process of the initially infected subpopulation is very erratic. The probability of an infection to persist is low, and the variation in the time to successful infection of the first initially uninfected subpopulation is quite large (figure 8). This is in agreement with well-known results, e.g. about infection persistence,[37,64–67] about population dynamics,[36,68–70] and about the persistence probability of a dominant mutant gene with a small selective advantage.[71] We also show that the infection process of other subpopulations follows a pattern of frequent extinction and rescue cycles, which have already been observed by other authors in metapopulations (Ref. 72, and Ref. 36, p. 148f). Even successfully infected subpopulations do not contribute heavily to the infection of other, still uninfected subpopulations (figure 8). During the infection process of a metapopulation, the initially infected subpopulation is the main contributor of infected cells. Conversely, this means that if the initially infected subpopulation dies out too soon, the infection is at risk.

4.3. *Caveats*

We now discuss some limitations of our study.

First, the spatial distribution of habitat patches for host cells of an IS infection is not known, and it may vary strongly for different landscapes. This lack of knowledge may be furthermore aggravated by our uncertainty about dispersal rates of prokaryotes over different distances. For example, a steeply decreasing dispersal rate function might reduce the infection speed of an IS much more in a landscape with strongly clumped habitat patches than in a landscape with uniformly distributed habitat patches. However, available dispersal data of prokaryotes in the wild[54] suggest that their dispersal rate function has a fat tail and does not decrease steeply with increasing distance, which reduces the dependence of the infection speed on the spatial distribution of habitat patches. In addition, because many ISs can move by HGT among different genera of prokaryote hosts[5] with different dispersal rate functions, some of which may well have a fat tail, the dependence of the infection speed on the spatial distribution of habitat patches may be further reduced. Moreover, we simulated a wide range of spatially structured and unstructured metapopulations (figure 5) and used different dispersal rate functions (figure 6). With the exception of the nearest neighbour rate function, we did not find any large differences in the time to full infection, which suggests that on the spatial scales we consider, spatiality has only a limited impact on the time to full infection.

Second, there is considerable uncertainty in the parameters that govern local

processes within a subpopulation or within host cells, i.e. the HGT rate h, the fitness cost and benefit s_d and s_b of an IS, the IS conversion rates c_d and c_b, and the IS excision rate e. To compensate for this uncertainty, we conducted simulations using a range of values for these parameters.

Third, we assumed that local and global process parameters stay constant over time and space, at least in relation to the rate of cell division, and we also assumed that no externally induced catastrophic events take place (e.g. no subpopulation extinction caused by habitat destruction). Our interest lies mainly in the influence of spatiality on infection dynamics.

Fourth, for reasons of simulation feasibility, we had to restrict the geographical size of all metapopulations to a diameter between 100 and 400 km. Our results are therefore only valid for a geographical region of about that size.

4.4. *Conclusion*

Despite the limitations mentioned above, our results allow us to make the following qualitative assertions. First, an IS infection is an erratic process during its early phase, both in a single population and in a metapopulation consisting of several subpopulations. Second, the initially infected subpopulation is the driving force of the IS infection in a metapopulation, so that the success of a metapopulation infection mainly depends on the success of the infection in this subpopulation. The extinction probability in the initially infected subpopulation is high, and even if the IS infection in this subpopulation succeeds, it takes many failed attempts until another subpopulation is successfully infected. Third, for the dispersal rates and the spatial scales we study, spatiality and dispersal do not strongly reduce infection speed, in contrast to local processes within a subpopulation or within a host cell.

5. Authors contributions

MB, ADB, and AW designed the study. MB wrote the program and conducted the simulations. MB and ADB performed the analytical calculations. MB and AW wrote the paper.

6. Acknowledgements

MB and AW acknowledge support through Swiss National Science Foundation grant 31003A_146137. AW acknowledges support by ERC Advanced Grant 739874, by Swiss National Science Foundation grant 31003A_172887, as well as by the Research Priority Program Evolution in Action at the University of Zürich. ADB was supported in part by Australian Research Council Grants nos. DP120102728 and DP120102398.

Appendix A. Transposition, excision and HGT rates

Table 3 shows a summary of reported IS transposition, IS excision and HGT rates that are used as a reference for our model parameters.

Table 3. Transposition, excision, and HGT rates reported by different authors.

Event		Rates	Sources
Transposition	Conservative	10^{-7} – 10^{-4}	Refs. 5, 8, 45
Excision		10^{-10} – 10^{-6}	Refs. 8, 45
HGT	Transformation	10^{-6} – 10^{-3}	Ref. 73
	Transduction	10^{-8}	Ref. 74
	Conjugation	10^{-6} – 10^{-5}	Ref. 75

Note: Rates have been converted to numbers of events per cell or IS and generation.

Appendix B. The tau-leaping algorithm

Originally, the tau-leaping algorithm was developed to approximately simulate the dynamics of a chemically reacting system,[48] as an extension of the exact but computationally expensive Doob-Gillespie algorithm.[76] We use the tau-leaping algorithm to simulate the infection dynamics inside a well-mixed subpopulation. The formulation of this algorithm below will therefore be adapted to accommodate such a situation.

The basic algorithm works as follows. Assume that a population consists of different types of individuals (e.g. uninfected, detrimentally infected or beneficially infected cells). Everything that can happen to an individual is called an event. There are different classes of events, i.e. one event class may be a specific change of type (e.g. detrimental infection of an uninfected cell), and death or propagation of an individual of a specific type (e.g. cell division of an uninfected cell). The rates with which those different event classes happen may depend on the numbers of individuals of any type (e.g. density-dependent death or infection by HGT). For an event of any class, we know how it will affect the numbers of individuals of different types (e.g. the detrimental infection of an uninfected cell will decrease the number of uninfected cells by one and increase the number of detrimentally infected cells by one). Assume that we know the numbers of individuals of all different types at a specific time t_0. We can then calculate the rates with which all classes of events will happen at that time. Based on these rates, we can calculate the maximal time span τ during which all the events that are expected to happen do not change the numbers of individuals of different types by too much, and therefore do not change the rates of all different event classes by too much either. Having calculated that time span τ (from which the algorithm got its name), we then determine for each

event class the number of events ocurring during τ by drawing a random number from a Poisson distribution with a mean equal to the product of τ and the event rate in that class at time t_0. Finally, we use all the events of all classes that have been determined to calculate the numbers of individuals of all types at time $t_0 + \tau$. One τ-leap has then been executed.[48]

The basic algorithm described in the last paragraph works well to simulate the population dynamics as long as none of the numbers of individuals of any type are small. If at any time the number of individuals of a type gets small, there is a risk that during the next τ-leap the events that have been determined to happen would by chance require to decrease the number of individuals of that type to negative values. To avoid such situations, one distinguishes between noncritical and critical event classes at time t_0. Noncritical are those classes which will presumably not reduce any numbers of individuals to negative values during the next time step. (This is usually implemented by imposing a threshold value on the number of events of a class that can happen without reducing the numbers of individuals of any involved types to negative values.) For those event classes, we calculate τ_{noncrit} using the basic algorithm. We call an event class critical, if during the next τ-leap, events of that class might result in negative numbers of individuals for at least one type. To avoid this, we determine the time span τ_{crit} until the next event of any of the critical event classes will happen by drawing a random number from an exponential distribution with mean $1/r$ where r is the sum of the rates of all critical event classes. The idea is to make sure that at most one event of any critical class happens during the next τ-leap. We therefore choose $\tau = \min(\tau_{\mathrm{crit}}, \tau_{\mathrm{noncrit}})$ as the length of the next τ-leap. The numbers of events of noncritical classes that occur during the next τ-leap are determined using the basic algorithm. If $\tau = \tau_{\mathrm{noncrit}}$, no event of any critical class will happen during the τ-leap, and if $\tau = \tau_{\mathrm{crit}}$, exactly one event of any critical class will happen. We determine the class to which this event belongs by drawing a random number from a distribution with probabilities proportional to the rates of all critical event classes. Finally, we use all events of all classes (noncritical and perhaps critical) to calculate the numbers of individuals of all types at time $t_0 + \tau$.[49]

Appendix C. The branching process model

Our multi-type, continuous-time Markov branching process model (Ref. 77, p. 199ff) consists of two types: detrimentally and beneficially infected cells (panel B in figure 1). At any time, four different events can happen to a cell: cell division, cell death (including IS excision), IS conversion, and HGT. Observe that we do not include uninfected cells in our model, and therefore IS excision from a detrimentally or beneficially infected cell is treated as cell death. In addition, since we assume that HGT always leads to detrimentally infected cells, an HGT event in a detrimentally infected cell can be treated the same way as cell division. We assume the waiting time to a cell's next event to have an exponential distribution with mean $1/a_d$ for

detrimentally infected cells and $1/a_b$ for beneficially infected cells, where

$$a_d = 2 + s_d + e + c_b + h$$
$$a_b = 2 + s_b + e + c_d + h$$

At the time of an event, the probabilities of the four different event types are given by the following values:

cell type	cell div.	cell death	IS conversion	HGT
detrimentally inf.	$\frac{1+s_d+h}{a_d}$	$\frac{1+e}{a_d}$	$\frac{c_b}{a_d}$	0
beneficially inf.	$\frac{1+s_b}{a_b}$	$\frac{1+e}{a_b}$	$\frac{c_d}{a_b}$	$\frac{h}{a_b}$

Based on the event probabilities at the time of an event, we obtain the probability generating functions

$$g_d(z_d, z_b) = \frac{1 + s_d + h}{a_d} z_d^2 + \frac{1 + e}{a_d} + \frac{c_b}{a_d} z_b$$
$$g_b(z_d, z_b) = \frac{1 + s_b}{a_b} z_b^2 + \frac{1 + e}{a_b} + \frac{c_d}{a_b} z_d + \frac{h}{a_b} z_d z_b$$

From the probability generating functions, we derive the infinitesimal generating functions

$$\begin{aligned} \tilde{g}_d(z_d, z_b) &= (1 + s_d + h)\, z_d^2 - (2 + s_d + e + c_b + h)\, z_d + c_b z_b + 1 + e \\ \tilde{g}_b(z_d, z_b) &= (1 + s_b)\, z_b^2 - (2 + s_b + e + c_d + h)\, z_b + c_d z_d + h z_d z_b + 1 + e \end{aligned} \tag{C.2}$$

and the infinitesimal generator

$$A = \begin{pmatrix} s_d - e - c_b + h & c_b \\ c_d + h & s_b - e - c_d \end{pmatrix}. \tag{C.3}$$

The eigenvalue λ_0 of A with the largest real part is itself real. If $\lambda_0 < 0$, $\lambda_0 = 0$ or $\lambda_0 > 0$, the branching process is called subcritical, critical, or supercritical, respectively.

Subcritical and critical branching processes become extinct with certainty. Supercritical branching processes have a non-zero probability of survival. The survival probability of an IS infection starting with one beneficially infected cell is given by $p_{\text{surv}} = 1 - p_{\text{ext}}$, where p_{ext} is the infection's extinction probability. The extinction probability of an IS infection starting with one beneficially infected cell is the component q_b of the smallest root $\mathbf{q} = (q_d\ q_b)^T$ of the infinitesimal generating function $\tilde{\mathbf{g}}(\mathbf{z})$ (equation (C.2)) in the interval $[\mathbf{0}, \mathbf{1}]$ (Ref. 77, p. 205).

If the branching process is supercritical, there exist positive right and left eigenvectors $\mathbf{u} = (u_d\ u_b)^T$ and $\mathbf{v} = (v_d\ v_b)^T$ of the infinitesimal generator A (equation (C.3)), which can be scaled so that $u_d + u_b = 1$ and $u_d v_d + u_b v_b = 1$. In the following we always assume that this scaling has been done.

The following holds in a supercritical, irreducible, multi-type branching process with finite 2nd moment, as described by our model (Ref. 78, p. 157f):

(1) There exists a random variable $W_k^m(t) := \frac{Z_k^m(t)}{v_k e^{\lambda_0 t}} \overset{t\to\infty}{\longrightarrow} W^m$ for any $m \in \{d, b\}$ and $k \in \{d, b\}$, where $Z_k^m(t)$ is the number of cells of type k at time t, starting with one cell of type m at time $t = 0$.
(2) the characteristic function $\varphi^m(x) = \mathbb{E}\left(e^{iW^m x}\right)$ of W^m, where $i = \sqrt{-1}$, obeys the system of ordinary differential equations $\frac{d\varphi^m(x)}{dx} = \frac{\tilde{g}^m(\varphi^1(x),\ldots,\varphi^l(x))}{\lambda_0 x}$.

After numerically solving the ordinary differential equation system for the characteristic functions $\varphi^m(x)$, we use the Fourier inversion theorem to reconstruct the probability density f^b of the random variable W^b from its characteristic function φ^b as $f^b(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi^b(x)dx$. We then calculate the number $Z_k^b(t)$ of detrimentally ($k = d$) and beneficially ($k = b$) infected cells at time t (big enough for the approximation to hold) in a population that has been infected with one cell containing one IS in its genome as $Z_k^b(t) = v_k e^{\lambda_0 t} W^b$. The total size of the population of infected cells is $Z(t) := Z_d^b(t) + Z_b^b(t) = e^{\lambda_0 t} W^b (v_d + v_b)$. Therefore, the time T_N to a threshold of N infected cells, for N sufficiently large, is

$$T_N = \frac{1}{\lambda_0}\left[\ln(N) - \ln(W^b) - \ln(v_d + v_b)\right],$$

and the median time to this threshold of infected cells is

$$T_{N,\text{med}} = \frac{1}{\lambda_0}\left[\ln(N) - \ln(W^b_{\text{med}}) - \ln(v_d + v_b)\right],$$

where W^b_{med} is the median of the random variable W^b, which can be computed using the density f^b of W^b.

The observation mentioned above that $\frac{Z_k^m(t)}{v_k e^{\lambda_0 t}} \overset{t\to\infty}{\longrightarrow} W^m$ for $m \in \{d, b\}$ and $k \in \{d, b\}$ means that the proportion of detrimentally and beneficially infected cells at time t (big enough for the approximation to hold) is given by $v_d/(v_d + v_b)$ and $v_b/(v_d + v_b)$, respectively.

In an earlier paper we showed that a purely detrimental IS infection can only persist if HGT can overcome an IS's fitness cost, i.e. if $h > |s_d|$.[20] Taking into account that the conversion rates c_b and c_d are small, the persistence of an IS infection for $h \leq |s_d|$ therefore depends on the persistence of the population of beneficially infected cells. To calculate the survival probability of such an IS infection with $h \leq |s_d|$, we can thus simplify the multi-type model to a birth-and-death process model, where the state of the model corresponds to the number of beneficially infected cells, and process state 0 is considered to be absorbing, meaning that the population of beneficially infected cells has become extinct. The birth and death rates per infected cell are $1 + s_d$ and $1 + e \approx 1$, respectively. According to a result by Kendall,[79] the probability of the birth-and-death process being in state 0 (i.e. the population of beneficially infected cells having died out) at time t then is

$$P_0(t) = \frac{1 - e^{-s_b t}}{1 + s_b - e^{-s_b t}}.$$

The survival probability of an IS infection with $h \leq |s_d|$ can therefore be approximated by

$$p_{\text{surv}} = 1 - \lim_{t\to\infty} P_0(t) = 1 - \frac{1}{1+s_b} \approx s_b \text{ for small } s_b, \tag{C.7}$$

where we have linearised around $s_b = 0$ in the last step.

This agrees with a result by Haldane,[71] who showed that a dominant mutant gene with a small selective advantage s, so that the expected number of offspring is $1 + s$, persists with a probability of about $2s$ in a random mating population. Considering the small fitness advantage s_b of a beneficially infected cell, the expected number of offspring of such a cell is $[2 \cdot (b + s_b) + 0 \cdot d] \,/\, (b + s_b + d) \approx 1 + s_b/2$ (linearisation for small s_b), where $b = d = 1$ are the division and death rates, respectively, of an uninfected cell in a subpopulation at carrying capacity. In our model, the selective advantage assumed by Haldane is therefore $s = s_b/2$, and the probability of a beneficially infected cell to persist and spread through the subpopulation is $2s = s_b$.

Appendix D. The spreading of an IS infection inside a metapopulation is irregular

Figure 9 shows for four different dispersal rate functions snapshots of the IS infection process, when for the first time at least one third of all subpopulations of a metapopulation are infected. The figure shows that whether an IS infection spreads regularly or irregularly depends strongly on the dispersal rate function. For example, infections based on power law dispersal rate functions (panels *p* and *sp* in figure 9) proceed irregularly, with many isolated infected subpopulations.

Appendix E. The spatial size of a metapopulation has only a moderate effect on the time to full infection

To assess the influence of a metapopulation's size on the IS infection dynamics, we also conducted 5000 simulations for metapopulations with default parameters in circular regions with a radius of 50 km and of 200 km instead of the default radius of 100 km. Using a hexagonal lattice with the same distance of 12.5 km between two neighbouring subpopulations as in our default simulations, these two regions contained 61 and 931 subpopulations, respectively. Figure 10 shows the time to full infection for the two new metapopulations and the original metapopulation with 241 subpopulations. While there is no noticeable difference in the time to full infection between a radius of 50 km and a radius of 100 km, the time to full infection in a circular region of 200 km radius is somewhat longer than in the other two regions. However, the difference in the median times to full infection is not large. For the circular region with radius 200 km, the median time increases by only 3.3% in comparison to the region with radius 100 km.

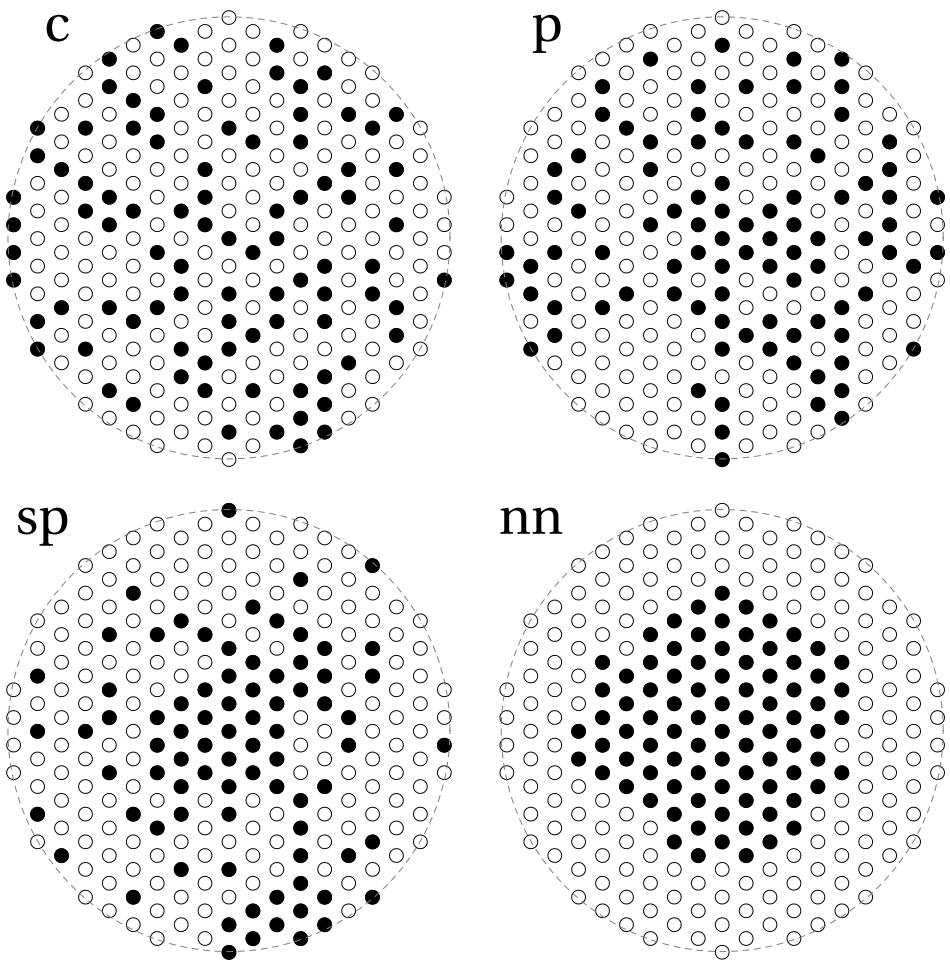

Fig. 9. *The spreading of an IS infection inside a metapopulation is irregular.* The four panels show, for four representative simulations, snapshots of infected (closed circles) and uninfected (open circles) subpopulations at the first time when at least one third of all subpopulations in the metapopulation are infected. Each of the four simulations is representative of all simulations with default parameters from one of the following four dispersal rate functions: constant (c), default power law (p), steep power law (sp), and nearest neighbour (nn). For each dispersal rate function, the representative simulation is chosen so that the simulation's time to full infection is the one closest to the median time to full infection of all simulations with this dispersal rate function.

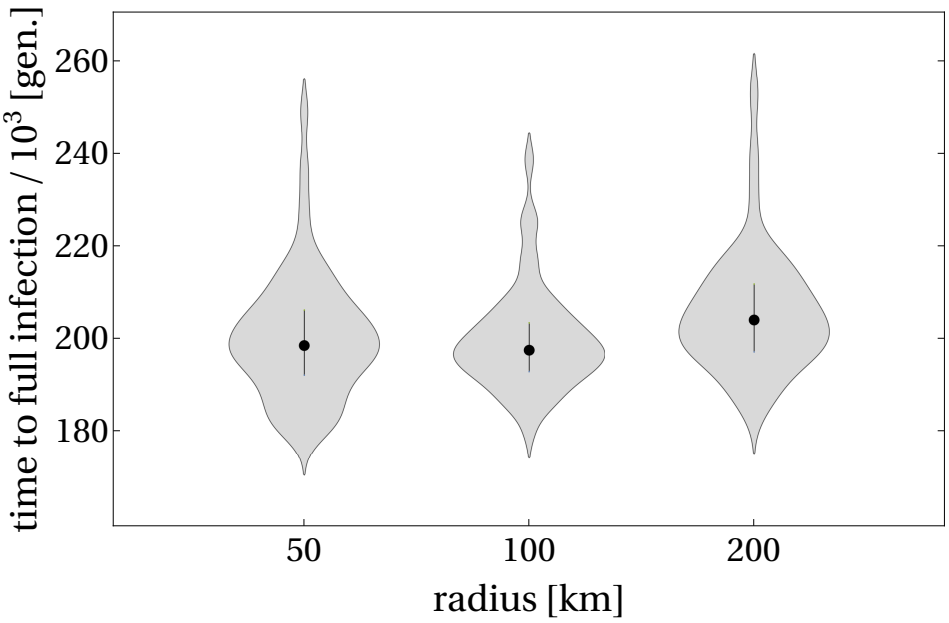

Fig. 10. *The spatial extent of a metapopulation influences IS infection dynamics only weakly.* Violin plot showing a kernel density estimate (shaded region), the median (dot) and the first and third quartile (endpoint of whiskers) of the time to full infection for three different radii of the circular region which encloses a metapopulation. Number of observations (left to right): 61, 49, and 47 simulations that led to fully infected metapopulations. Due to excessive simulation times, only 4807 out of 5000 simulations finished when using a radius of 200 km.

1. Barbara McClintock. The origin and behaviour of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, 36(6):344–355, 1950.
2. Donal A. Hickey. Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics*, 101:519–531, 1982.
3. Brian Charlesworth, Paul Sniegowski, and Wolfgang Stephan. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, 371:215–219, 1994.
4. John F. Y. Brookfield. The ecology of the genome – mobile DNA elements and their hosts. *Nature Reviews Genetics*, 6(2):128–136, 2005.
5. Michael Chandler and Jacques Mahillon. Insertion sequences revisited. In Nancy L. Craig, Robert Craigie, Martin Gellert, and Alan M. Lambowitz, editors, *Mobile DNA II*, pages 305–366. American Society for Microbiology, Washington, D.C., 2002.
6. Howard Ochman, Jeffrey G. Lawrence, and Eduardo A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405:299–304, 2000.
7. David J. Galas and Michael Chandler. Bacterial insertion sequences. In Douglas E. Berg and Martha M. Howe, editors, *Mobile DNA*, pages 109–162. American Society for Microbiology, Washington, D.C., 1989.
8. Nancy Kleckner. Transposon Tn10. In Douglas E. Berg and Martha M. Howe, editors, *Mobile DNA*, pages 227–268. American Society for Microbiology, Washington, D.C., 1989.
9. Dominique Schneider and Richard E. Lenski. Dynamics of insertion sequence elements during experimental evolution of bacteria. *Research in Microbiology*, 155:319–327, 2004.
10. Patricia Siguier, Edith Gourbeyre, and Mick Chandler. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews*, 38:865–891, 2014.
11. Magdalene So and Brian J. McCarthy. Nucleotide sequence of the bacterial transposon TN1681 encoding a heat-stable (ST) toxin and its identification in enterotoxigenic Escherichia coli strains. *Proceedings of the National Academy of Sciences of the United States of America*, 77(7):4011–4015, 1980.
12. Douglas E. Berg. Transposon Tn5. In Douglas E. Berg and Martha M. Howe, editors, *Mobile DNA*, pages 185–210. American Society for Microbiology, Washington, D.C., 1989.
13. Eva M. Top and Dirk Springael. The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Current Opinion in Biotechnology*, 14:262–269, 2003.
14. Michel Blot. Transposable elements and adaptation of host bacteria. *Genetica*, 93(1-3):5–12, 1994.
15. James A. Shapiro. Transposable elements as the key to a 21st century view of evolution. *Genetica*, 107:171–179, 1999.
16. Richard Dawkins. *The selfish gene*. Oxford University Press, 1976.
17. W. Ford Doolittle and Carmen Sapienza. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284:601–603, 1980.
18. Leslie Eleazer Orgel and Francis Harry Compton Crick. Selfish DNA: the ultimate parasite. *Nature*, 284:604–607, 1980.
19. Sergey V. Nuzhdin. Sure facts, speculations, and open questions about the evolution of transposable element copy number. *Genetica*, 107:129–137, 1999.
20. Manuel Bichsel, A. D. Barbour, and Andreas Wagner. The early phase of a bacterial insertion sequence infection. *Theoretical Population Biology*, 78:278–288, 2010.
21. Manuel Bichsel, A. D. Barbour, and Andreas Wagner. Estimating the fitness effect of an insertion sequence. *Mathematical Biology*, 66:95–114, 2013.

22. Andreas Wagner. Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Molecular Biology and Evolution*, 23(4):723–733, 2006.
23. Hadi Quesneville and Dominique Anxolabéhère. Dynamics of transposable elements in metapopulations: A model of *P* element invasion in *Drosophila*. *Theoretical Population Biologiy*, 54(2):175–193, 1998.
24. Grégory Deceliere, Sandrine Charles, and Christian Biémont. The dynamics of transposable elements in structured populations. *Genetics*, 169(1):467–474, 2005.
25. Stanley A. Sawyer and Daniel L. Hartl. Distribution of transposable elements in prokaryotes. *Theoretical Population Biology*, 30(1):1–16, 1986.
26. Christopher J. Basten and Michael E. Moody. A branching-process model for the evolution of transposable elements incorporating selection. *Journal of Mathematical Biology*, 29:743–761, 1991.
27. Lourens Gerhard Marinus Baas Becking. *Geobiologie of inleiding tot de milieukunde*. W. P. Van Stockum & Zoon, The Hague, the Netherlands, 1934.
28. Noah Fierer. Microbial biogeography: patterns in microbial diversity across space and time. In K. Zengler, editor, *Accessing Uncultivated Microorganisms: from the Environment to Organisms and Genomes and Back*, pages 95–115. ASM Press, Washington DC, 2008.
29. Hamish McCallum, Drew Harvell, and Andy Dobson. Rates of spread of marine pathogens. *Ecology Letters*, 6:1062–1067, 2003.
30. S. M. Burrows, T. Butler, P. Jöckel, H. Tost, A. Kerkweg, U. Pöschl, and M. G. Lawrence. Bacteria in the global atmosphere – part 2: Modeling of emissions and transport between different ecosystems. *Atmospheric Chemistry and Physics*, 9:9281–9297, 2009.
31. Christina A. Kellog and Dale W. Griffin. Aerobiology and the global transport of desert dust. *TRENDS in Ecology and Evolution*, 21(11):638–644, 2006.
32. Dominic Mao and Dennis Grogan. Genomic evidence of rapid, global-scale gene flow in a *Sulfolobus* species. *The ISME Journal*, 6:1613–1616, 2012.
33. Kelle C. Freel, Anna Edlund, and Paul R. Jensen. Microdiversity and evidence for high dispersal rates in the marine actinomycete '*Salinispora pacifica*'. *Environmental Microbiology*, 14(2):480–493, 2012.
34. R. Thane Papke and David M. Ward. The importance of physical isolation to microbial diversification. *FEMS Microbiology Ecology*, 48:293–303, 2004.
35. David Tilman and Peter Kareiva, editors. *Spatial ecology: the role of space in population dynamics and interspecific interactions*. Princeton University Press, New Jersey, 1997.
36. Ilkka Hanski. *Metapopulation ecology*. Oxford University Press, New York, 1999.
37. Matt J. Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, New Jersey, 2008.
38. Mark Shaffer. Minimum viable populations: coping with uncertainty. In Michael E. Soulé, editor, *Viable populations for conservation*, pages 69–86. Cambridge University Press, Cambridge, 1987.
39. Russell Lande. Risks of population extinction from demographic and environmental stochasticity and random catastrophes. *The American Naturalist*, 142(6):911–927, 1993.
40. John M. Halley and Yoh Iwasa. Extinction rate of a population under both demographic and environmental stochasticity. *Theoretical Population Biology*, 53:1–15, 1998.
41. Otso Ovaskainen and Baruch Meerson. Stochastic models of population extinction.

Trends in Ecology and Evolution, 25(11):643–652, 2010.
42. Marie Touchon and Eduardo P. C. Rocha. Causes of insertion sequences abundance in prokaryotic genomes. *Molecular Biology and Evolution*, 24(4):969–981, 2007.
43. Henry Rodriguez, Elizabeth T. Snow, Uppoor Bhat, and Edward L. Loechler. An *Escherichia coli* plasmid-based, mutational system in which *supF* mutants are selectable: Insertion elements dominate the spontaneous spectra. *Mutation Research*, 270(2):219–231, 1992.
44. Michael Lynch. *The origins of genome architecture*. Sinauer Associates, Inc., Sunderland, 2007.
45. Ana Sousa, Catarina Bourgard, Lindi M. Wahl, and Isabel Gordo. Rates of transposition in *Escherichia coli*. *Biology Letters*, 9, 2013.
46. Laura S. Frost, Raphael Leplae, Anne O. Summers, and Ariane Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, 2005.
47. Michael Lynch and John S. Conery. The origins of genome complexity. *Science*, 302:1401–1404, 2003.
48. Daniel T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115(4):1716–1733, 2001.
49. Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics*, 124(044109):1–11, 2006.
50. Andrew D. Barbour, Kais Hamza, Haia Kaspi, and Fima C. Klebaner. Escape from the boundary in Markov population processes. *Advances in Applied Probability*, 47(4):1190–1211, 2015.
51. David M. Olson, Eric Dinerstein, Eric D. Wikramanayake, Neil D. Burgess, George V.Ñ. Powell, Emma C. Underwood, Jennifer A. D'Amico, Illanga Itoua, Holly E. Strand, John C. Morrison, Colby J. Loucks, Thomas F. Allnutt, Taylor H. Ricketts, Yumiko Kura, , John F. Lamoreux, Wesley W. Wettengel, Prashant Hedao, and Kenneth R. Kassem. Terrestrial ecoregions of the world: a new map of life on earth. *BioScience*, 51(11):933–938, 2001.
52. R. V. O'Neill, J. R. Krummel, R. H. Gardner, G. Sugihara, B. Jackson, D. L. DeAngelis, B. T. Milne, M. G. Turner, B. Zygmunt, S. W. Christensen, V. H. Dale, and R. L. Graham. Indices of landscape pattern. *Landscape Ecology*, 1(3):153–162, 1988.
53. Habin Li and James F. Reynolds. A new contagion index to quantify spatial patterns of landscapes. *Landscape Ecology*, 8(3):155–162, 1993.
54. Michael S. Roberts and Frederick M. Cohan. Recombination and migration rates in natural populations of *Bacillus subtilis* and *Bacillus mojavensis*. *Evolution*, 49(6):1081–1094, 1995.
55. James M. Bullock and Ralph T. Clarke. Long distance seed dispersal by wind: measuring and modelling the tail of the curve. *Oecologia*, 124:506–521, 2000.
56. M. W. Shaw, T. D. Harwood, M. J. Wilkinson, and L. Elliott. Assembling spatially explicit landscape models of pollen and spore dispersal by wind for risk assessment. *Proceedings of the Royal Society B*, 273:1705–1713, 2006.
57. Theodore E. Harris. *The theory of branching processes*. Dover Publications Inc., New York, 1989.
58. Patsy Haccou, Peter Jagers, and Vladimir A. Vatutin. *Branching Processes: Variation, Growth, and Extinction of Populations*. Cambridge University Press, New York, 2005.
59. Montgomery Slatkin and Wayne P. Maddison. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, 123:603–6013, 1989.
60. Ran Nathan, Frank M. Schurr, Orr Spiegel, Ofer Steinitz, Ana Trakhtenbrot, and Asaf

Tsoar. Mechanisms of long-distance seed dispersal. *Trends in Ecology and Evolution*, 23(11):638–647, 2008.
61. Mark Kot, Mark A. Lewis, and P. van den Driessche. Dispersal data and the spread of invading organisms. *Ecology*, 77(7):2027–2042, 1996.
62. Denis Mollison. The rate of spatial propagation of simple epidemics. In Lucien M. Le Cam, Jerzy Neyman, and Elizabeth L. Scott, editors, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pages 579–614. University of California Press, Berkeley, California, 1972.
63. M. A. Lewis and S. Pacala. Modeling and analysis of stochastic invasion processes. *Journal of Mathematical Biology*, 41:387–429, 2000.
64. M. S. Bartlett. Deterministic and stochastic models for recurrent epidemics. In Jerzy Neyman, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 81–109. University of California Press, Berkeley and Los Angeles, California, 1956.
65. B. T. Grenfell. Chance and chaos in measles dynamics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(2):383–398, 1992.
66. F. Ellis McKenzie, Gerry F. Killeen, John C. Beier, and William H. Bossert. Seasonality, parasite diversity, and local extinctions in *Plasmodium falciparum* malaria. *Ecology*, 82(10):2673–2681, 2001.
67. N. Dexter. Stochastic models of foot and mouth disease in feral pigs in the Australian semi-arid rangelands. *Journal of Applied Ecology*, 40(2):293–306, 2003.
68. Robert H. MacArthur and Edward O. Wilson. *Island Biogeography*. Princeton University Press, Princeton, NJ, 1967.
69. John M. Drake and David M. Lodge. Allee effects, propagule pressure and the probability of establishment: risk analysis for biological invasions. *Biological Invasions*, 9:365–375, 2006.
70. Tim M. Blackburn, Julie L. Lockwood, and Phillip Cassey. The influence of numbers on invasion success. *Molecular Ecology*, 24:1942–1953, 2015.
71. John Burdon Sanderson Haldane. A mathematical theory of natural and artificial selection, part V: selection and mutation. *Proceedings of the Cambridge Philosophical Society*, 23:838–844, 1927.
72. James H. Brown and Astrid Kodric-Brown. Turnover rates in insular biogeography: Effect of immigration on extinction. *Ecology*, 58(2):445–449, 1977.
73. Haydn G. Williams, Martin J. Day, John C. Fry, and Gregory J. Stewart. Natural transformation in river epilithon. *Applied and Environmental Microbiology*, 62(8):2994–2998, 1996.
74. Sunny C. Jiang and John H. Paul. Gene transfer by transduction in the marine environment. *Applied and Environmental Microbiology*, 64(8):2780–2787, 1998.
75. Cecilia Dahlberg, Maria Bergström, and Malte Hermansson. In situ detection of high levels of horizontal plasmid transfer in marine bacterial communities. *Applied and Environmental Microbiology*, 64(7):2670–2675, 1998.
76. Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.
77. Krishna B. Athreya and Peter E. Ney. *Branching Processes*. Springer Verlag, Berlin, 1972.
78. Boris A. Sewastjanow. *Verzweigungsprozesse*. Akademie Verlag, Berlin, 1975.
79. David G. Kendall. On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics*, 19(1):1–15, 1948.